DOCUMENT RESUME

ED 106 770 95 CS 001 800

AUTHOR Manelis, Leon
TITLE "The American Heritage Word Frequency Book" and Its Relation to the Communication Skills Lexicon. Technical Note No. 2-72-38.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
REPORT NO SWRL-TN-2-72-38
PUB DATE Sep 72
NOTE 22p.

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Communication Skills; *Educational Research; Elementary Secondary Education; Measurement Instruments; Readability; Reading Instruction; *Reading Skills; *Textbooks; *Word Frequency; Word Recognition
IDENTIFIERS *American Heritage Word Frequency Book

ABSTRACT

This paper reviews "The American Heritage Word Frequency Book" by John B. Carroll, Peter Davis, and Barry Richman (CDR). CDR is a word frequency count derived from school books for children in grades 3-9. Included in this review are an explanation of the development of CDR and a guide to its practical use. The guide can be used independently of the rest of the paper. Comments are also on the usefulness of CDR for compilation of the Communication Skills Lexicon. Finally, various word frequency counts and frequency measures are compared and discussed in terms of their utility for assigning frequency values to words in the lexicon. (Author)

SOUTHWEST REGIONAL LABORATORY TECHNICAL NOTE

DATE: September 15, 1972

NO: TN 2-72-38

THE AMERICAN HERITAGE WORD FREQUENCY BOOK AND ITS RELATION TO THE COMMUNICATION SKILLS LEXICON

Leon Manelis

ABSTRACT

This paper reviews *The American Heritage word frequency book* by John B. Carroll, Peter Davies, and Barry Richman (CDR). CDR is a word frequency count derived from school books for children in grades 3-9. Included in this review are an explanation of the development of CDR and a guide to its practical use. The guide can be used independently of the rest of the paper. Comments are also made on the usefulness of CDR for compilation of the Communication Skills Lexicon. Finally, various word frequency counts and frequency measures are compared and discussed in terms of their utility for assigning frequency values to words in the Lexicon.

THE AMERICAN HERITAGE WORD FREQUENCY BOOK AND ITS RELATION TO THE COMMUNICATION SKILLS LEXICON

Leon Manelis

The American Heritage word frequency book by Carroll, Davies, and Richman (CDR) is a word frequency count derived from school books for children in grades 3-9. The corpus from which it was developed was used as a citation base for the *American Heritage school dictionary*. The total number of tokens in the corpus was about 5 million, and the number of types, about 87,000. A word type was defined as a string of graphic characters bounded at left and right by spaces; graphic characters included letters, numerals, internal punctuation (hyphen and apostrophe) and some mathematical symbols. This means that a base word and its inflected variants all have separate entries. Upper and lower case letters were distinguished, thus producing separate entries for words that were capitalized in the corpus and those that were not. Capitalization was not coded, however, simply for words at the beginning of sentences.

The word types are ordered in two ways: alphabetically and by rank according to frequency. In the alphabetical list, CDR gives an elaborate array of information for each type. The frequency of occurrence in the corpus is given as a single number (F). This number is also broken down into eight grade levels (3-9 and ungraded). F is again broken down into seventeen subject areas. The grade level and subject area assignments were determined by a consensus of school personnel who recommended the sources from which the corpus was drawn.

In addition to the simple frequency, three derived statistics are given for each word type. D is a measure of the dispersion of frequency across subject areas. U is an estimate of the "true" frequency in a theoretically infinite corpus rather than the finite corpus actually sampled. The estimate is made on the basis of the dispersion of sampled frequency across subject areas; a word type that is evenly distributed has a higher U value than a type whose frequency is concentrated in one area. SFI is a logarithmic transformation of U. CDR suggests that once understood, SFI is a simple and convenient way of indicating the probability of occurrence of a given type. In addition to this information for each word type, CDR gives a statistical analysis of the corpus as a whole and an extensive set of frequency distribution graphs.

SAMPLING OF THE SOURCES FROM WHICH THE CORPUS WAS DRAWN

A survey of school systems in the United States was conducted in November and December of 1969. Schools surveyed were mostly public systems with large enrollments; Roman Catholic and private systems were also included. For each type of system, an attempt was made to maintain an even geographic distribution. Questionnaires were sent to the highest administrators in the systems, and they often delegated completion of the questionnaires to other personnel. The respondents were asked to list "the textbooks, individual study and practice materials, library books, and other reading matter most commonly used in your grades 3 through 9." They listed titles according to subject area and according to grade level as determined by use in their own school systems. Each

title was assigned to a single grade and subject based on the modal choice of recommendations; this introduced a bias in favor of the lower grades. If there was no modal subject, a title was assigned to the first subject in an established order; the order emphasized basal or standard curriculum areas (reading was first, religion was last). About 6,000 titles were recommended. Of these, about 1,000 were selected to form a corpus of the desired size. In the sample of 1,000, the same proportions of titles in subject areas and grade levels were maintained as in the set of 6,000. Within this constraint, the most frequently recommended titles were selected. Thus, the final sample of 1,000 titles accurately represented the original survey.

Ten thousand samples were taken from these sources; each sample included 500 words of running text. For each grade and subject, a constant number of samples was drawn from all the sources, regardless of their lengths. (The number was based on the proportion of the total number of recommendations made for the grade and subject.) From each source, that number of 500-word samples was drawn at uniform intervals beginning on the first page. Thus, for a given grade and subject, the same number of words was drawn from short books as from long books.

A LIGHTNING GUIDE TO THE PRACTICAL USE OF CDR

In CDR, the most concise information on its use is given in the "Guide to the Alphabetical List," which is on pages 1-4. These are the most valuable pages in the book. The following is an even more abbreviated initiation to CDR, but it is intended to provide sufficient

information for the occasional user.

Opening CDR to a randomly selected page of the Alphabetical List, you see an imposing array of numbers and words. Do not panic. The words are listed at the extreme left and at the bottom of the page. First consider the words at the left. These had frequencies of at least 2 occurrences in the corpus of 5 million words (tokens). The exact number of occurrences is given in the column headed F, which is immediately to the right of the words. Most of the other numbers on the page belong to one of two breakdowns of F. In the columns headed Gr 3, Gr 4, ..., Gr 9, UnGr, F is broken down according to grade level. A 5 under Gr 3, for example, means that the word you're interested in occurred five times in samples of text from books that are typically used in third grade. (UnGr indicates books assigned to an ungraded category.) The sum of all the numbers in the grade columns is equal to F. In the columns headed Read, Eng & Gr, Comp, and so on, F is broken down according to subject area. The sum of all the numbers in these columns also equals F. The headings are fairly obvious abbreviations, but an explanation of them is given in the table on page 2. The classification of a book in a subject area was done on the basis of the survey described above. A 4 under Art, for example, means that the word you're interested in occurred four times in samples drawn from books used in art instruction.

A strong warning should be issued in interpreting the grade level and subject area breakdowns. The number of tokens sampled varies across grade and subject categories. The number of tokens represented in the category of reading, for example, was over one million; in religion, less than 5,000. Thus the frequency breakdowns cannot be compared across

categories without reference to distributions of tokens sampled. These distributions are given in the first column of the table on page xxxvii.

Three columns in the Alphabetical List give other statistics: D, U, and SFI. D is a measure of the dispersion of tokens across subject areas. It ranges from 0 to 1, with lower values indicating a concentration of tokens in a few subject areas and higher values, a more even distribution.

U is an estimate of the "true" frequency of a word type in a theoretically infinite corpus. For a given F, words with a low value of D have a lower U value than words with a high value of D. U is scaled in terms of frequency per million, and it assumes fractional values less than as well as greater than one.

SFI (standard frequency index) is a logarithmic transformation of U. It is theoretically justified in that its distribution is approximately normal (p. xxxi). From a practical standpoint it can be interpreted in terms of handy frequency categories. A word type with an SFI value of 40 would be expected to occur once in a million tokens; with a value of 50, ten times in a million tokens; a value of 60, one hundred times; and so on.

The simple frequencies and the derived statistics differ from each other in an important way. Whereas F or its components can validly be summed across word types to yield a value for a class of words, this is not the case for D, U, and SFI. This is important to keep in mind if you want to pool the base form of a word with its inflected variants. Details for combining the statistics are given on page 3, but for D, U, and SFI, the procedures are probably too complicated for practical use.

It should be pointed out, however, that simply adding U values does give an approximation to the more complicated computation.

Now consider the words at the bottom of a page in the Alphabetical List. All of these occurred only once in the sample of 5 million tokens. The grade levels and subject areas in which they were found are coded by numbers and letters next to the words. The numbers indicate grade level (X, however, means ungraded), and the letters indicate subject area according to the key at the top of each page. For all these words, D = 0, and values of U and SFI are given for each subject area in the table on page 2; U and SFI are probably unreliable for a frequency of one, however.

A few words might be said about the physical layout of CDR. The size of the type is small, and there are many numbers on each page. It is advisable to use a marker in order to keep one's place and to block off some of the visual array. Another unfortunate aspect of the layout is the separate list of words at the bottom of each page in the Alphabetical List. Although all of the entries on a page are within the alphabetical guide words at the top, it may be necessary to look in two places to find a given word. Both of these physical problems are manageable, however.

The Rank List is simpler than the Alphabetical List; breakdowns by grade level and subject area are not included. All of the words in the corpus are listed in order of their values on U and SFI. (The two variables are equivalent for the purpose of ranking.) Tied words are counted separately in the ranking. After every one hundred items, the rank number is given. To find an exact rank, it is necessary to count the number of items before or after a marked entry.

TABLE 1

THE SCOPE OF FOUR MAJOR WORD FREQUENCY COUNTS

	Number of Types	Words Not Found in a Sample of 250
Carroll, Davies, and Richman (1971)	86,741[1]	13
Kučera and Francis (1967)	50,406[1]	44
Kučera and Francis, including inflected types[2]		36
Thorndike and Lorge (1944)	19,440[3]	40
Thorndike and Lorge Juvenile Count	---[4]	45
Rinsland (1945)	14,571[5]	63

[1]Graphic types, including numbers and inflected words.

[2]A base form was considered present if an inflected variant of it was listed.

[3]The main listing contains this number of words. The remaining 10,560 words in the Thorndike-Lorge count have frequencies of less than one per million and are in two other lists. These were not consulted. Entries in the Thorndike-Lorge count are generally base forms. Inflected variants are usually included in the frequency value for a base form, but there are some separate listings as well.

[4]There are fewer than 19,440 words in the Juvenile Count (which is included in the main listing), but Thorndike and Lorge (1944) do not give the exact number.

[5]Includes inflected types.

THE USEFULNESS OF CDR FOR THE MOD 3 LEXICON

The primary strength of CDR is its extensiveness. It includes 87,000 word types, more than any other word frequency count. Table 1 shows the number of types in four major counts, including the number for CDR, which exceeds the others by at least 36,000. Table 1 also shows the number of words not found in the various counts out of a sample of 250. The sample comprised 55 words randomly selected from the Entry List of the Mod 3 Entry Lexicon (Rhode, 1972a) and 195 words randomly selected from a preliminary version of the Mod 3 General Lexicon (August, 1972). its greater coverage, there were far fewer words not found in CDR than in the other sources.

On the basis of its scope, CDR would be a useful source for future work on lexicon. The Rank List would be especially useful in selecting words above a criterion frequency. (The frequency distributions can be consulted to find the number of words above a given frequency.) Even at the present advanced stage of development of the Mod 3 General Lexicon, CDR might still supplement the current work. A frequency criterion might be established, taking into account the number of words to be considered. The resulting list of words would be edited according to existing inclusion-exclusion criteria (Cronnell, 1971; Rhode, 1972b). The remaining words would then be checked against the current Lexicon. Any words that might be added to the Lexicon should then be checked for grade level in the Alphabetical List. CDR words are taken from sources used in grades 3-9, but the Lexicon is designed for K-6. To exclude words from CDR that represent grades 7-9, the grade level distributions of F should be

consulted for each word. Words concentrated in the upper grades would not be added to the Lexicon. Twenty-six percent of the corpus of tokens is from sources used in grades 7-9 (p. xxxvii). Accordingly, if more than about one-fourth of the occurrences of a word are concentrated in these grades, the word should probably not be added to the Lexicon.

Although CDR could be useful for the General Lexicon, it is probably not suitable as a source for the Technical Lexicon. Cronnell and Rhode (1972) found it inadequate in a limited comparison with a set of music terms derived from two music texts also used by CDR. This is understandable in light of the sampling procedure used in compiling CDR; there was no attempt to exhaustively list all the words characteristic of a subject area. CDR may be useful in another way, however, rather than as a source. Cronnell (1971) stated that it may be difficult to decide whether a word should be assigned to the General or to the Technical Lexicon. The subject area distributions of F for each word could help to determine the assignment. If a word is used specifically in a particular subject, its occurrences in the corpus should be concentrated in that subject. The grade level distribution might have a similar use in determining grade placement, although only grades 3-6 in CDR would be relevant for the Lexicon. If the subject area or grade level breakdowns are used in this way, care should be taken to weight the component frequencies according to the distribution of tokens in the entire corpus.

CORRELATIONAL COMPARISONS OF WORD FREQUENCY COUNTS

CDR and the other three major frequency counts listed in Table 1

(Kučera and Francis, 1967; Rinsland, 1945; Thorndike and Lorge, 1944) are candidates for use in assigning frequency values for all words in the Communication Skills Lexicon. As an indication of the extent to which these sources differ, correlations were computed for the sample of 250 words described above. This sample comprised 55 Entry words and 195 General words, representing about the same proportion of Entry to General words as in the complete Lexicon. Within this constraint, selection of the words was completely random. The appendix to the paper lists the words selected. In assigning frequency values, only graphic types were looked up in Rinsland (R), Thorndike-Lorge (TL), and CDR. In Kučera-Francis, inflected variants were considered as well. One variable (KF) was the frequency of the graphic type alone. For another variable (KFI), if a word was the base form of an adjective, noun, or verb, inflected forms of the word were also looked up and if found, their frequencies were added to the frequency of the base form. (Only the affixes -(e)s, -(e)d, -ing, -er, and -est were considered, as specified by Cronnell (1971)). From the Thorndike-Lorge count, the "G" values printed in boldface were used. For high frequency words, which Thorndike and Lorge (1944) mark only with A or AA, numerical values were obtained by summing across the four components of G (the Thorndike, Lorge, Juvenile, and Semantic Counts). In addition to this variable (TL), the Juvenile Count alone was also used (TLJ). (The two lists of very low frequency words in the Thorndike-Lorge book were not consulted.) In CDR, both U and F were used (CU and CF, respectively).

All of these variables were correlated with each other. In addition, the logs of all the variables were computed, and these values were also correlated with each other. The reason for transforming the scores was the nature of the distribution of word frequencies. They tend to concentrate at the lower values and be dispersed at the higher values. This type of distribution inflates the correlation coefficient. The log transformation spreads the scores more evenly. (The transformation may also be theoretically justified by the fact that word frequencies have a lognormal distribution--log frequencies are normally distributed--and the Pearson correlation coefficient assumes a normal distribution for each variable.)

The resulting correlations are shown in Table 2. These figures are based on all 250 words. Because of the substantial number of zero values (representing words not present in a count), two other sets of correlations were also computed. One set excluded all (93) words that had a zero value on any variable. In the other set, a word with a zero on a given variable was excluded only from the correlations into which that variable entered. The pattern of results for these two sets was the same as that about to be discussed for the correlations based on all 250 words.

The correlation between CU and CF was very high: .997 for the simple scores and .964 for the logs. This suggests that there is little difference between the two measures.

The correlation between KF and KFI was also very high: .993 for the simple scores and .972 for the logs. For many of the words, of course, there was no difference between KF and KFI because the words

TABLE 2

CORRELATIONS OF WORD FREQUENCIES

Untransformed Frequencies

	R	KF	KFI	TL	TLJ	CU
R						
KF	.749					
KFI	.736	.993				
TL	.878	.947	.943			
TLJ	.518	.616	.650	.693		
CU	.754	.914	.906	.876	.651	
CF	.760	.927	.920	.894	.660	.997

Log Transformations

	R	KF	KFI	TL	TLJ	CU
R						
KF	.733					
KFI	.714	.972				
TL	.788	.834	.868			
TLJ	.651	.668	.708	.898		
CU	.860	.877	.854	.844	.684	
CF	.855	.870	.856	.837	.700	.964

taken from the Lexicon either did not have inflected forms or else were already inflected. But there were differences between KF and KFI for 119 of the 250 words. The very high correlation in spite of these differences suggests that for the purpose of obtaining frequencies, no information is gained by adding inflected variants to the base form of a word.

The correlations of TLJ with the other variables were the lowest of the whole set. (The one exception was the correlation of log TLJ with log TL.) This may be due to an oddity of the Juvenile Count. For the most frequent words an exact frequency is not given. Instead, they are simply marked to indicate a frequency of at least 1,000 in 4.5 million tokens. These words were assigned a frequency of 1,000 in the correlations. Restriction of the range of frequency in this way may have reduced the correlations. (The restriction of the range may also account for the unexpected effect of the log transformation. In all six cases, the correlations of TLJ with the other variables were greater for the logs than for the simple scores. Of the remaining 15 correlations, 13 were smaller for the logs, as expected.)

Aside from the correlations with TLJ, the correlations among the Kučera-Francis, Thorndike-Lorge, and Carroll counts were relatively high and not much different from each other. This pattern of results provides no basis for differentiating among the three sources. Correlations of the three with the Rinsland count were slightly lower, suggesting that there may be something different about the Rinsland count. One can speculate that the difference is in the sources used in compiling the counts; Rinsland is based on children's writings,

and the others are based on published materials.

Considering the correlations alone, one cannot make any strong recommendations about which source to use for assigning frequencies to Lexicon words. There is a slight implication that the choice would be between Rinsland on the one hand and Kucera-Francis, Thorndike-Lorge, and CDR on the other hand. Other considerations should enter into the decision, however. CDR has much to recommend it above the other sources. It is the most recent and, of primary significance, it is the most extensive. It also presents a great deal of information that can be practically useful, including frequency distributions and the subject area and grade level breakdowns. In particular, the subject area breakdowns could be of help in determining whether a word belongs in the General or Technical Lexicon. Another potentially useful feature is the SFI measure, a log transformation of simple frequency. This could be convenient for establishing frequency categories. As discussed above, the log transformation smoothes out the positively skewed distribution of frequencies; in effect, it compresses the scale for high frequencies and expands it for low values. It should be borne in mind, however, that SFI values cannot be summed across word types. The particular values to be used as cutoffs for the frequency categories would ideally be established by considering the distribution of values in the entire Lexicon. Then categories could be formed so as to include equal numbers of words. A priori cutoffs would make equal-sized categories less likely.

The correlations do clearly suggest that in CDR there is little if any difference between the U and F measures when considering a large

sample of words. There is also the suggestion, based on the very high correlation of KF with KFI, that there is little if any difference in using graphic types alone as oppposed to including the variants of a base form. This might be attributable to a relatively small increment due to including the inflected variants, as compared with the differences between distinct words. The conclusion about inflected words is somewhat counter-intuitive. It seems that a base form is a meager representation of all the variants that share something of its meaning. Although the correlation between KF and KFI is convincingly high, it may be reassuring to try out both types of frequency assignments on a subset of words when the assignments are to be used in the Mod 3 Lexicon itself. If frequency values are to be used in sequencing, for example, it might be observed whether the two types of assignment imply different sequences.

SUMMARY AND CONCLUSIONS

The development of CDR was described, and a guide to its practical use was provided. The guide explains the information presented in the main listings of CDR, and it can stand by itself. Comments were made on the usefulness of CDR for compilation of the Communication Skills Lexicon. Finally, correlations of various word frequency counts and frequency measures were presented and discussed.

The following conclusions can be drawn from this report:

1. CDR is the most extensive word frequency count available.
2. CDR provides a great deal of statistical information, some of which may be practically useful.

a. The subject area breakdowns may help in determining whether a word should be assigned to the General or Technical Lexicon.

b. SFI may be convenient for establishing frequency categories.

3. In a large sample of words, there is little difference between using the U or F measures from CDR.

4. In assigning frequency values, there is probably little difference between using the base form of a word alone or including its inflected variants.

5. CDR is recommended above the other major word frequency counts as a source for frequency values in the Communication Skills Lexicon.

REFERENCES

Carroll, J.B., Davies, P., & Richman, B. *The American Heritage word frequency book.* Boston: Houghton Mifflin, 1971.

Cronnell, B. Inclusion-exclusion criteria for the Model 3 communication skills lexicon. Technical Note No. TN-2-71-38, 1971. Southwest Regional Laboratory, Los Alamitos, California.

Cronnell, B., & Rhode, M. A tryout of procedures for developing a technical lexicon. Technical Note No. TN-2-72-17, 1972. Southwest Regional Laboratory, Los Alamitos, California.

Kučera, H., & Francis, W.N. *Computational analysis of present-day American English.* Providence, Rhode Island: Brown University Press, 1967.

Rhode, M. Specification of the Model 3 Entry Lexicon. Technical Note No. TN-2-72-08, 1972. Southwest Regional Laboratory, Los Alamitos, California. (a)

Rhode, M. Specification of the Model 3 General Lexicon. Technical Note No. TN-2-72-30, 1972. Southwest Regional Laboratory, Los Alamitos, California. (b)

Rinsland, H.D. *A basic vocabulary of elementary school children.* New York: MacMillan, 1945.

Thorndike, E.L., & Lorge, I. *The teacher's word book of 30,000 words.* New York: Teachers College, Columbia University, 1944.

APPENDIX

WORD SAMPLE USED FOR CORRELATIONAL STUDY

General

dimly
reassure
fleece
mosquito
bellow
teaspoon
rigging
bazaar
wistful
neglectful
formula
clergyman
squirm
disposal
repent
chart
boring
hoof
sensation
equation
intended
bole
reflect
outboard
eyelid
caste
contribute
uncertainty
bye-bye
hostess
chrysanthemum
occupation
sling
baseman
imprisonment
lag
gloat
riches
alter
deal
par
mere
possessed

shoveling
pretense
aisle
bruise
heifer
laughter
limb
herb
prevailing
borax
whipping
swimmer
miracle
prison
hardening
independent
description
armchair
area
liberty
hereafter
freeman
echo
halo
sage
pizza
petroleum
servant
photographer
toil
doom
delegate
context
nursing
homeless
candidate
boom
excellence
char
automotive
crackle
newborn
stagnant

judge
elbow
whip
parka
balk
gremlin
cosmic
tugboat
gardening
evident
recite
moment
milkweed
sheepherder
pop-top
interest
assure
lumberjack
hijack
accent
bedspread
bullfight
jumper
lecture
paratrooper
hark
acquaintance
bounty
livestock
tariff
ladle
research
reckless
flier
forgotten
repeat
cafe
peak
classroom
tow
harpoon
seeker
dying

jogging
outer
dialogue
mestizo
festival
bloom
exhaust
chilled
register
frowning
pedal
clank
cavemen
copperhead
soggy
famous
altar
teenibopper
labor
shotgun
pelt
northland
retreat
diminish
crash-land
establish
severe
stratosphere
plight
dusty
grateful
auxiliary
trainer
sneaker
charger
expensive
magnolia
legislation
assemble
hither
mercy
farewell
crackajack
advanced
inspect
bin
wriggle
haze
historian
powerhouse

racing
run-through
insane
gracefully
buccaneer
ward
solemnly
hobby
meadow
sandy
chock
mom
livery
flyer
hearty
snorkel

Entry

many
joke
lying
hood
chair
sprinkler
good
desk
stone
sauce
pour
bring
else
electric
sled
page
scrub
so
little
clothespin
class
million
set
uncover
prize
grandfather
push
Thanksgiving
rainbow
license

him
sell
mattress
bird
choo
elevator
crowd
grass
cost
pill
my
ride
canoe
bossy
cabinet
good-bye
turkey
straight
skeleton
sailor
feeling
jar
drip
yes
tight

DISTRIBUTION

1 Baker

1 Berger

1 Bessemer

5 Cronnell

1 Follettie

1 Koehler

1 Legum

15 Library

1 Linda Lester

1 Manelis

1 O'Hare

1 Rhode

1 Rudegeair

1 Rundle

1 Russell

1 Schutz

1 Terry